A Lingüística para o processamento das línguas

Eric Laporte

A Lingüística para o processamento das línguas

Éric Laporte

Laboratório de Informática do Instituto Gaspard-Monge
Universidade de Marne-la-Vallée, França
www-igm.univ-mlv.fr/~laporte

Departamento de Letras
PUC-RJ

O processamento das línguas é uma área que, embora não seja nova, é percebida como tal, talvez por causa de um enraizamento insuficiente nos dois campos da lingüística e da informática. Nessa apresentação geral da área, esclareceremos primeiro o que é o processamento das línguas, e depois examinaremos as relações com a lingüística e com a teoria do léxico-gramática. Concluiremos sobre as contribuições da informática no processamento das línguas, e também sobre a adequação do termo "lingüística computacional" que é muito usado como sinônimo de "processamento das línguas".

Colocação do problema

O que é o processamento das línguas ? Nesse termo, a palavra *processamento* designa a atuação de um sistema num computador. Trata-se de uma seqüência de operações automáticas, pela qual um conjunto de dados iniciais, as entradas, são processadas, e um outro conjunto de dados, os resultados, são produzidos. No caso de aplicações informáticas tradicionais, as entradas e os resultados são freqüentemente dados numéricos, ou pelo menos estritamente codificados, por exemplo: tabelas de valores financeiros ou estatísticos. Neste parágrafo, usamos a palavra *entradas* na acepção informática (em inglês, *input*), não se trata ainda de *entradas* lexicais, a noção mais fundamental da lexicologia (*lexical entries*) !

O processamento das línguas consiste, de fato, em qualquer processamento de textos em línguas naturais[1]. Quer as entradas do sistema, quer os resultados, são textos escritos ou falados, por exemplo, em português. Hoje em dia, muitos usuários de computadores estão familiarizados com vários produtos comerciais, cuja função é processar textos escritos:

- Os editores de textos permitem a aquisição e o arquivamento do texto dos documentos e do enriquecimento tipográfico. Possuem funções adicionais: imprimir, verificar a ortografia…
- Os sistemas de busca de páginas na Web processam periodicamente o conteúdo textual dos *sites* do mundo inteiro, e permitem achar uma seleção de páginas relacionadas com um assunto determinado.
- Os sistemas de ajuda à tradução traduzem textos de uma língua para outra.

[1] Línguas naturais são o objeto de estudo habitual da lingüística, em oposição às linguagens formais ou artificiais, que são meios de comunicação com o computador ou objetos matemáticos.

Esses três exemplos de sistemas informáticos estão disponíveis para um amplo público, e é fácil verificar que, embora sejam úteis, seu desempenho ainda não é satisfatório. Os melhores editores de textos apontam erros em palavras corretas, propõem correções erradas, e deixam de detectar certos tipos de erros ortográficos. Os sistemas de busca na Web selecionam, às vezes, dezenas de páginas sem qualquer relação com o assunto pesquisado pelo usuário, mesmo que este expresse seu objetivo de forma suficientemente precisa. Até os textos produzidos pelos melhores sistemas de ajuda à tradução necessitam uma releitura por tradutores humanos, por causa dos erros de tradução que, aliás, tornam os resultados da tradução automática quase um gênero literário cômico.

É preciso entender se os defeitos dos sistemas de processamento existentes são inerentes ao problema, ou se resultam de dificuldades que poderão ser superadas. Logo no início da história da informática, aproximadamente há cinqüenta anos atrás, pesquisadores começaram a trabalhar rumo ao desenvolvimento de vários tipos de processamento automático de textos: a criptografia e a tradução automática foram os primeiros, a análise sintática se seguiu. Durante esses cinqüenta anos, tanto a lingüística como a informática desenvolveram vias de pesquisa de modo quase independente. Trata-se de dois mundos, desde a origem, estranhos um ao outro. De um lado, o dos lingüistas, procedente dos filólogos, de outro, o dos engenheiros. Dois mundos que ainda estão aprendendo a se conhecer mutuamente e a colaborar eficazmente.

A via da lingüística parecia a opção mais racional, dado que o material a ser processado consistia em textos, portanto um material de natureza lingüística[2]. A obra do lingüista Zellig Harris e a teoria do léxico-gramática do lingüista Maurice Gross foram contribuições decisivas para encaminhar essa via e torná-la factível. Este é o assunto principal do nosso artigo.

A via da informática, porém, sempre foi mais explorada. No período recente, o aumento das potencialidades técnicas dos computadores foi rápido, e enormes quantidades de textos se tornaram disponíveis em suporte eletrônico. Numerosos sistemas de processamento de textos foram elaborados, muitas vezes em apenas alguns meses, com aplicação de métodos e aproximações matemáticos, e com integração de poucos dados lingüísticos, ou, em alguns casos, sem integração de qualquer informação lingüística. Nesses sistemas, a falta de dados lingüísticos é só parcialmente compensada pela efetuação de computações. Isso é um elemento de explicação do desempenho, ainda passível de melhoria, dos produtos disponíveis. A observação, a formalização e a integração de dados lingüísticos necessita um ritmo de elaboração mais lento, mas com certeza possibilitará progressos substanciais no desempenho dos sistemas.

A herança de Harris

O lingüista Z. Harris (1952) desenvolveu uma teoria lingüística original, pela sua orientação para a forma diretamente observável da língua. Em sua visão, a fonte básica do conhecimento lingüístico é a observação de fatos inteiramente superficiais : a aceitabilidade de frases, por exemplo. A aceitabilidade de:

[2] Pelo mesmo raciocínio, no caso de uma aplicação informática relacionada, por exemplo, ao escoamento da água, o estudo das equações de mecânica dos fluidos seria considerado objetivamente como o primeiro passo rumo à realização de sistemas.

(1) *O presidente aumentou os salários*

e a inaceitabilidade de:

* *O aumentou presidente os salários*

são constatadas por julgamentos diretos, embora subjetivos, de um falante da língua. O asterisco marca que uma seqüência é inaceitável como frase. A existência de uma relação de sinonimia entre (1) e a seguinte frase é outro exemplo de fato diretamente observável:

(2) *Os salários foram aumentados pelo presidente*

As definições são rigorosas e econômicas. A metodologia empiricista de Harris evita a criação e a manipulação de construções abstratas e complexas, de regras, de níveis, que não sejam estritamente necessários para descrever ou formalizar os fatos observáveis ou para simplificar esta formalização. A limitação à mera descrição combinatória da língua possibilita a construção de gramáticas coerentes. Esta metodologia não é aplicável a todos os campos da lingüística: por exemplo, no caso da diacronia, a língua não tem forma diretamente observável. Porém, quando é aplicável, orienta o lingüista à procura de um apoio formal às suas intuições, por exemplo intuições de explicações.

Uma das ferramentas teóricas mais úteis neste quadro metodológico é a noção harrissiana de transformação sintática (Harris, 1964, 1968). Existem dois tipos principais de transformações:

- as transformações unárias, por exemplo a passiva, que se aplicam a uma frase elementar, e
- as transformações binárias, como a coordenação e a subordinação de frases, que combinam duas estruturas em outra estrutura.

A passiva e as pronominalizações são transformações unárias bem conhecidas. Conservam o sentido das frases às quais se aplicam, portanto permitem a constituição de classes de equivalência semântica, como no exemplo seguinte:

(1) *O presidente aumentou os salários*
= *Aumentou os salários*
= *O presidente aumentou-os*
= *Aumentou-os*

= (2) *Os salários foram aumentados pelo presidente*
= *Foram aumentados pelo presidente*
= *Os salários foram aumentados por ele*
= *Foram aumentados por ele*

O símbolo "=" marca o fato de que duas frases são ligadas por uma transformação sintática. A última frase da série acima resulta, primeiro, da aplicação da passiva, e depois, de duas pronominalizações: uma pronominalização do sujeito e uma do complemento em *por*. Muitas outras transformações podem ser descritas, por exemplo a média:

(1) = (3) *Os salários aumentaram*

As transformações têm um caráter regular, isto é, aplicam-se de forma idêntica a numerosas frases. Assim, a relação entre (1) e (3) é observada também entre (4) e (5):

(4) *O João apagou a luz*
= (5) *A luz apagou*

Todavia, esta regularidade está limitada pelas restrições de aplicação das transformações. Por exemplo, a transformação média não se aplica a (6), o resultado da aplicação sendo inaceitável:

(6) *O João tirou o documento da gaveta*
** O documento tirou da gaveta*

As transformações binárias operam sobre duas estruturas. O resultado é uma estrutura complexa, por exemplo :

- uma coordenação (*O presidente aumentou os salários, mas a medida foi cancelada*);
- uma subordinação adverbial (*A luz apagou porque a lâmpada queimou*);
- uma construção relativa (*O presidente aumentou os salários, que não mudaram em vários anos*).

O léxico-gramática

Essas ferramentas metodológicas e teóricas abriram um programa de pesquisa bastante amplo. O lingüista Maurice Gross definiu a teoria do léxico-gramática, segundo a qual os objetivos da lingüística necessitam de uma descrição lexical efetiva de grande porte por falantes nativos (1975). Nenhuma teoria lingüística possibilitara a descrição combinatória completa de uma língua. Diante deste fato surpreendente, era natural duvidar se tal descrição ainda fazia parte dos objetivos dessas teorias. Também era natural afirmar que a lingüística descritiva, e singularmente a descrição lexical, permaneciam um objetivo prioritário.

Com certeza, a busca de explicações de fatos lingüísticos é interessante, mas já uma descrição dos mesmos fatos seria útil. Certos fatos lingüísticos não têm outra explicação a não ser contingências históricas, que podem ser reconstituídas ou não: onde não existe explicação, a descrição do fato é indispensável. E onde uma explicação pode ser achada, o conhecimento do fato permanece uma etapa prévia à explicação. Portanto, a concepção do léxico-gramático decorreu do seguinte raciocínio: na ausência de um amplo programa de descrição lexical, os objetivos da lingüística moderna não passariam de uma absurda campanha de explicação de fatos desconhecidos.

Um grupo de lingüistas do LADL[3] descreveu importantes segmentos da gramática do francês, com um objetivo de exaustão do ponto de vista lexical. O método de

[3] Laboratoire d'automatique documentaire et linguistique, Université Paris 7 – Centre national de la recherche scientifique.

investigação (Gross, 1990, 1994) implicou o exame individual de 12.000 verbos, do tipo de (1), de mais de 20.000 verbos compostos, do tipo de :

(7) *O João soltou os cachorros*

bem como de 500 transformações sintáticas, que foram sistematicamente confrontadas às entradas lexicais estudadas. Como vimos nos exemplos (4) a (6), as condições de aplicação das transformações incluem condições lexicais, isto é, uma transformação é geralmente aplicável a muitas frases e inaplicável a muitas outras. Uma descrição morfológica, mais simples do que esta descrição sintática, abrange aproximadamente 800.000 palavras simples e 130.000 palavras compostas.

O léxico-gramática respeita uma distinção estrita entre dicionário e corpus. O dicionário, construído para a descrição mais exaustiva possível do léxico, leva em conta todas as propriedades gramaticais cuja formulação precisa depende das entradas gramaticais, por exemplo, a maioria das transformações sintáticas. Até a aplicação de transformações raras a entradas lexicais raras faz parte de seu objetivo. Um corpus de textos, pelo contrário, por vasto que seja, tem o estatuto de amostra de formas da língua. É muito útil como fonte de exemplos, mas não ensina nada a respeito de formas que não constarem nele.

Segundo um dos princípios fundamentais do léxico-gramática, a unidade mínima de sentido é a frase elementar, constituída por um predicado (um verbo no exemplo (1), mas pode ser um predicado nominal ou adjetivo) com seu sujeito e seus complementos essenciais. Esta opção teórica resulta dos dois fatos seguintes:

- o estudo de uma palavra isolada priva o descritor da possibilidade de avaliar aceitabilidades, já que o julgamento de aceitabilidade se aplica a frases;
- numa frase elementar, o contexto tira muitas vezes a ambigüidade da palavra isolada.

Assim, representando o sujeito e os complementos essenciais pelos símbolos N_0, N_1, N_2..., as unidades mínimas de sentido de uma língua podem aparecer como fórmulas:

N_0 *aumentar* N_1
N_0 *apagar* N_1
N_0 *tirar* N_1 *de* N_2
N_0 *soltar os cachorros*

No caso deste último exemplo, o complemento direto é fixo: se for substituído por outro substantivo, perde-se o sentido da expressão; não faria sentido representar uma forma lingüística fixa pelo símbolo N_1, que representa uma distribuição de complementos possíveis.

Os laboratórios e grupos de pesquisa que empreenderam a construção de um léxico-gramático constituem uma rede chamada RELEX[4]. A maior parte deles são europeus. O português (Eleutério et al., 1995; Ranchhod et al., 1999) e o inglês (Machonis, 1988) são duas das línguas sendo descritas neste quadro metodológico e teórico.

[4] www.ladl.jussieu.fr/Relex/RELEX.html.

Conclusão

A lingüística para o processamento das línguas possui duas características essenciais: a enormidade da tarefa a ser cumprida, e a importância das aplicações técnicas.

É interessante colocar a questão da contribuição da informática no programa científico e técnico que apresentamos, e no qual a pertinência da lingüística é claramente reconhecida. O armazenamento e a manutenção dos dados lingüísticos, mesmo antes de realizar qualquer aplicação informática, já requerem o computador. Para confrontar o dicionário com um texto, e associar às palavras do texto as informações lingüísticas do dicionário, precisamos de ferramentas de análise lexical (Silberztein, 1997) que são também elementos essenciais das futuras aplicações. Os problemas resolvidos por ferramentas informáticas de manutenção de dicionários e análise lexical são problemas clássicos: armazenamento de dados volumosos em pouco espaço, acesso rápido a dados volumosos, agrupamento de análises hipotéticas paralelas, reconhecimento de formas numa seqüência linear… mas não se trata de problemas de computação no sentido normal. É por esta razão que não incluímos, no título deste artigo, o termo de *lingüística computacional.*

Referências

Eleutério, Samuel, Elisabete Ranchhod, Helena Freire, Jorge Baptista, 1995. A System of Electronic Dictionaries of Portuguese, *Lingvisticae Investigationes* XIX:1, Amsterdam: Benjamins, pp. 57-82.

Gross, Maurice. 1975. *Méthodes en syntaxe*, Paris: Hermann, 412 p.

Gross, Maurice. 1990. Sur la notion harrissienne de transformation et son application au français, *Langages* 99, Paris: Larousse, pp. 39-56.

Gross, Maurice. 1994. Constructing Lexicon-Grammars. *Computational Approaches to the Lexicon*, Atkins and Zampolli (eds.), Oxford University Press, pp. 213-263.

Harris, Zellig. 1952. Discourse Analysis, *Language* 28, Baltimore: Waverly Press, pp. 1-30.

Harris, Zellig. 1964. Elementary transformations, Philadelphie: University of Pennsylvania, *Transformations and Discourse Analysis Papers* 54. Reimpresso em *Papers on Syntax. Structural and Transformational Linguistics*, 1970, Henry Hiz (ed.), Reidel: Dordrecht, pp. 211-235.

Harris, Zellig. 1968. *Mathematical Structures of Language*, New York: Wiley, 230 p.

Machonis, Peter, 1988. Support verbs: An Analysis of *be Prep X* idioms, *The SECOL Review* 122, pp. 95-125.

Ranchhod, Elisabete, Cristina Mota, Jorge Baptista, 1999. A computational lexicon of Portuguese for automatic text parsing. *SIGLEX 99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL*, pp. 74-80.

Silberztein, Max. 1997. The lexical analysis of natural languages, *Finite-State Language Processing*, Roche and Schabès (eds.), Cambridge: MIT Press, pp. 175-203.